

# تخمین سریع ضرایب پیچش در هنجارسازی طول مجرای صوتی با استفاده از امتیاز به دست آمده از مدل سازی جنسیت

یاسر شکفته<sup>۱</sup>، محمد محسن گودرزی<sup>۲</sup>، حسن قلی پور<sup>۳</sup>، سید جهان شاه کبودیان<sup>۴</sup>،

فرشاد الماس گنج<sup>۵</sup>، شقایق رضا<sup>۶</sup> و ایمان صراف رضایی<sup>۷</sup>

<sup>۱</sup> و <sup>۲</sup> و <sup>۳</sup> گروه پردازش صوت و زبان طبیعی، پژوهشگاه توسعه فناوری های پیشرفته خواجه نصیرالدین طوسی، تهران، ایران  
<sup>۴</sup> گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه رازی، کرمانشاه، ایران  
<sup>۵</sup> دانشکده مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، تهران، ایران

## چکیده

یکی از مشکلات عمده سامانه های خودکار بازشناسی گفتار (ASR)، تنوع موجود در بین گوینده ها، کانال انتقال داده و محیط است که به علت وجود این تنوعات، کارایی این سامانه ها در شرایط کاربردی مختلف، به شدت تغییر می کند. مقاوم سازی سامانه های بازشناسی جهت مقابله با این تغییرات از جمله مسائل حال حاضر در حوزه بازشناسی گفتار است. از جمله عواملی که باعث کاهش کارایی سامانه ها می شود، تمایز مشخصات صوتی آواهای یکسان تولید شده از گوینده های مختلف است. یکی از عوامل اصلی این مشکل ناشی از تفاوت موجود در طول مجرای صوتی (VTL) بین گوینده های مختلف است. روش هنجارسازی طول مجرای صوتی (VTLN) از روش های رایج برای رفع این مشکل است که در آن برای هر گوینده یک ضریب پیچش فرکانسی تعیین می شود. در این مقاله روش متداول تعیین ضریب پیچش با رویکرد مبتنی بر جستجو در یک سامانه بازشناسی گفتار پیوسته فارسی مبتنی بر مدل مخفی مارکوف معرفی و مشکلات محاسباتی استفاده از این روش شرح داده شده است. در نهایت روشی مبتنی بر رگرسیون خطی از روی امتیاز محاسبه شده از مدل سازی جنسیت جهت تخمین ضرایب پیچش پیشنهاد شده است که منجر به کاهش قابل ملاحظه هزینه محاسباتی روش مبتنی بر جستجو می شود. علاوه بر این، نتایج آزمایش ها بر روی دادگان آزمون گفتار تلفنی محاوره ای، بیان گر بهبود ۵۴/۰ درصدی دقت تشخیص کلمه روش پیشنهادی نسبت به روش متداول مبتنی بر جستجو است.

واژگان کلیدی: بازشناسی گفتار، هنجارسازی طول مجرای صوتی، تشخیص جنسیت، رگرسیون خطی، ضریب پیچش فرکانسی.

## ۱- مقدمه

شرایط کاربردی به شدت تغییر می کند. بنابراین مقاوم سازی سامانه های بازشناسی جهت مقابله با این تغییرات از جمله مسائل حال حاضر حوزه بازشناسی گفتار است. از جمله منابعی که باعث کاهش کارایی سیستم ها می شود، تغییرات بین آواهای تولید شده از گوینده های مختلف است. این مسئله که تأثیر به سزایی در به دست آمدن ویژگی های متمایز از سیگنال صوتی آواهای یکسان دارد، نه تنها ناشی از تغییرات موجود در فیزیک دستگاه های تولید صوت و آناتومی متفاوت آنهاست، بلکه ناشی از عوامل دیگری همچون تغییرات زبانی بین گوینده های مختلف مانند نحوه تلفظ، لهجه، استرس و ... است. یکی از عوامل اصلی تغییرات بین گوینده های مختلف، ناشی از این است که گویندگان

گفتار، روشی طبیعی برای ارتباط میان انسان ها است و می تواند نقش یک واسط را بین کاربر انسانی و ماشین ایفا کند. سامانه های خودکار بازشناسی گفتار (ASR) با هدف تبدیل گفتار به متن به وجود آمده اند. ورودی این سامانه ها، سیگنال گفتار و خروجی آن ها متن متناظر با گفتار ورودی است. این سامانه ها کاربردهای گوناگونی از قبیل سامانه های خودکار تلفنی، سامانه های خودکار دیکته نویسی و غیره دارند. کارایی این سامانه ها به میزان قابل توجهی وابسته به هماهنگی میان دادگان تعلیم و آزمایش آنها است؛ لذا از مشکلات عمده موجود در این سامانه ها می توان به تنوعات موجود در بین گوینده ها، کانال انتقال داده و محیط اشاره کرد که به علت وجود این تغییرات، کارایی این سامانه ها در

تقریب خوبی معادل با اعمال یک تبدیل خطی بر روی ضرایب کپسترال نهایی است و از آن برای هم‌تراز کردن فورمنت‌های<sup>۶</sup> گفتار کودکان با بزرگسالان استفاده شده است. همچنین، در (مکدونوهت، ۲۰۰۴) این تبدیل خطی به صورت یک فیلتر تمام‌گذر<sup>۷</sup> پیاده‌سازی شده است.

مزیت اصلی استفاده از تبدیل خطی در این است که فرآیند استخراج ویژگی تغییر نخواهد کرد و کافی است تا تبدیل مذکور را بر روی ویژگی‌های موجود اعمال کرد. این مزیت به خصوص در سامانه‌هایی که عملیات استخراج ویژگی در سامانه کاربر انجام شده اما عملیات بازشناسی در یک سامانه دیگر (به‌عنوان مثال سرور مرکزی) انجام می‌شود بروز پیدا می‌کند. البته دقت این شیوه اندکی پایین‌تر از روش VTLN عادی است که در آن پیش فرکانس به‌طور مستقیم در حوزه فرکانس و یا بر روی فیلتربانک اعمال می‌شود (پانچاپاگسان، ۲۰۰۹). همچنین کارایی این شیوه در شرایط وجود نوفه کاهش پیدا می‌کند و لذا کوشش‌هایی برای بهبود آن صورت گرفته است (سنند، ۲۰۱۰).

از طرف دیگر، می‌توان از این تبدیل خطی، به جای اعمال بر روی ویژگی‌ها برای اصلاح مشخصه‌های مدل سامانه بازشناسی استفاده کرد. در این شیوه مشابه روش تطبیق به‌گزینه «حداکثر درست‌نمایی رگرسیون خطی<sup>۸</sup> (MLLR)» عمل می‌شود و تبدیل خطی مذکور بر روی بردارهای میانگین مدل‌های HMM به کار رفته در سامانه بازشناسی اعمال می‌شود (کلیز، ۱۹۹۸؛ دینگ، ۲۰۰۲؛ ایموری، ۲۰۰۱). بدین ترتیب، این روش‌ها را می‌توان به‌عنوان یک مورد خاص از MLLR تفسیر کرد. البته می‌توان روش VTLN عادی را در کنار روش تطبیق به‌گزینه MLLR به کار برد و نشان داده شده است که این دو روش مکمل یکدیگر نیز هستند (گیولیانی، ۲۰۰۶).

در مرجع (باباعلی، ۱۳۸۶) روش‌های متداول استخراج و اعمال ضریب پیش بر روی طیف گفتار جهت هنجارسازی اثر طول مسیر صوتی مورد بررسی و مقایسه قرار گرفته‌اند. در مرجع (عزیزی، ۱۳۹۱) یک سامانه بازشناسی کلمات مجزا بررسی شده است که هدف آن افزایش کارایی سامانه بازشناسی گفتار کودکان با استفاده از روش هنجارسازی طول مسیر صوتی است. این سامانه بازشناسی، برای استفاده در نرم‌افزار گفتاردرمانی ایجاد شده بود، به‌طوری که نرم‌افزار نهایی با استفاده از سامانه

مختلف دارای طول مجرای صوتی<sup>۱</sup> (VTL) متفاوتی هستند. این تفاوت باعث می‌شود صدای افراد مختلف از لحاظ شنیداری با هم متفاوت باشد. یک روش متداول در کاهش تنوعات گوینده، بهبود تحلیل بانک فیلترها است. در این روش که هنجارسازی طول مجرای صوتی<sup>۲</sup> (VTLN) نام دارد، هدف جبران‌سازی اختلاف طول مسیر صوتی بین گویندگان مختلف است. حذف تغییرات ناشی از طول لوله صوتی با اعمال پیش فرکانسی روی طیف سیگنال گفتار گوینده انجام‌پذیر است. مقالات متعددی (لی، ۱۹۹۸؛ وگمان، ۱۹۹۶؛ پیتز، ۲۰۰۵؛ ایدی، ۱۹۹۶) به پیاده‌سازی و بررسی عملکرد روش‌های مختلف پیش فرکانسی پرداخته‌اند. اگرچه تطبیق به‌گزینه VTLN موضوعی با قدمت حدود دو دهه در حوزه پردازش و بازشناسی گفتار است؛ اما به‌علت کارایی اثبات‌شده آن، همچنان به‌عنوان یکی از فرآیندهای مؤثر در سامانه‌های مطرح و به‌روز بازشناسی گفتار (سائون، ۲۰۱۵؛ میترا، ۲۰۱۴؛ سایناس، ۲۰۱۴) به کار می‌رود.

روش VTLN به‌طور معمول با مقیاس‌دهی<sup>۳</sup> محور فرکانس (وانگ، ۲۰۰۷؛ کویی، ۲۰۰۶؛ آسرو، ۱۹۹۱) و یا با شیف‌دادن فرکانس‌های مرکزی فیلترهای میان‌گذر فیلتربانک (لی، ۱۹۹۸) پیاده‌سازی می‌شود. هر دو روش را می‌توان با استفاده از ضریب پیش<sup>۴</sup> انجام داد که این ضریب برای هر گوینده باید جداگانه محاسبه شود.

در مرجع (پیتز، ۲۰۰۵) نشان داده شده است که پیش فرکانسی در حوزه فرکانس پیوسته معادل با یک انتقال خطی<sup>۵</sup> (LT) در حوزه کپسترال پیوسته است و از این رو می‌توان به جای مقیاس‌دهی محور فرکانس، تبدیل خطی متناظر را بر روی تبدیل کپسترال اعمال کرد. البته لازمه این کار این است که یک رابطه تحلیلی برای انتقال از حوزه فرکانس به حوزه کپسترال به‌دست آورد و این رابطه به‌ناچار در حوزه فرکانس پیوسته تعریف می‌شود. البته در مرجع (اومش، ۲۰۰۵) نشان داده شد که با فرض محدود کردن بازه فرکانسی مورد نظر به یک بازه کوچک، این تبدیل خطی در حوزه فرکانس گسسته نیز به‌صورت تقریبی برقرار خواهد بود. به‌طور مشابه، در (کویی، ۲۰۰۶) نشان داده شده که انتقال فیلترهای مثلثی فیلتربانک مل در ضرایب MFCC با

<sup>1</sup> Vocal tract length

<sup>2</sup> Vocal tract length normalization

<sup>3</sup> Scaling

<sup>4</sup> Warping factor

<sup>5</sup> Linear Transform

<sup>6</sup> Formants

<sup>7</sup> All-pass

<sup>8</sup> MLLR

۱۹۹۸). این کار باعث کاهش قابل توجه زمان محاسبات می شود اما همچنان این روش، از آنجا که مبتنی بر جستجو می باشد، کند است. البته در صورتی که از روش های مبتنی بر تبدیل خطی استفاده شود، می توان رابطه درست نمایی را تا حد ممکن ساده کرد و سپس از روش امید ریاضی - بیشینه سازی<sup>۳</sup> (EM) و یا حتی روش های مبتنی بر کاهش گرادینان، عملیات تخمین ضریب پیچش را انجام داد (وانگ، ۲۰۰۷). اما به دلیل پایین بودن کارایی روش های مبتنی بر تبدیل خطی که در بالا ذکر شد، این روش ها مقبولیت کمتری دارند. در مرجع (ویلینگ، ۱۹۹۹) برای کاهش هزینه محاسبات روشی مبتنی بر GMM<sup>۴</sup> برای تخمین سریع VTLN ارائه شده است. در این روش، تمامی مقادیر مورد نظر ضریب پیچش بر روی داده های آموزش اعمال می شود و برای هر مقدار یک مدل GMM آموزش داده می شود. سپس در مرحله آزمون، ضریب پیچش متناسب با آن مدل GMM که بیشترین درست نمایی را دارد، انتخاب می شود. مطابق نتایج گزارش شده، اگرچه این روش هزینه محاسباتی را به میزان چشم گیری کاهش می دهد؛ اما همچنان روشی مبتنی بر جستجو است و در هر جستجو نیاز به محاسبه درست نمایی یک مدل GMM دارد. همچنین این نتایج نشان می دهند که این روش منجر به کاهش دقت بازشناسی نسبت به روش (لی، ۱۹۹۸) می شود.

در مرجع (میمورا، ۲۰۱۱) روشی مبتنی بر جدول گویندگان برای تشخیص ضریب پیچش ارائه شده است. در این روش برای تعدادی گوینده مرجع ضریب پیچش محاسبه می شود؛ سپس برای داده های آزمایش با استفاده از روش  $\Delta BIC$  نزدیک ترین گوینده از مجموعه فوق مشخص شده و از ضریب پیچش مربوط به آن گوینده برای داده آزمایش استفاده می شود. ادعا شده که این روش منجر به کاهش قابل توجه هزینه محاسباتی می شود؛ در حالی که میزان کاهش دقت بازشناسی محسوس نخواهد بود.

در مقاله حاضر، هدف ما نیز کاهش قابل ملاحظه هزینه محاسباتی تخمین ضریب پیچش فرکانسی، بدون کاهش دقت بازشناسی، در روش های مبتنی بر مقیاس دهی محور فرکانس است. می توان نشان داد که ضرایب پیچش ارتباط مستقیمی با جنسیت گوینده دارند. از آنجا که تشخیص جنسیت گوینده می تواند با هزینه محاسباتی اندکی انجام شود، در این مقاله، یک روش جدید برای تخمین

بازشناسی، درست یا نادرست بودن تلفظ کودک را تشخیص دهد و تلاش می کند تا با استفاده از بازخوردهای مناسب، گفتار کودک بهبود یابد. در مرجع (تبریزی، ۱۳۸۸) نیز روش VTLN برای بهبود کارایی بازشناسی گفتار کودکان فارسی زبان پیاده سازی شده است. همچنین یوما (یوما، ۲۰۱۳) یک روش VTLN فضای ویژگی ارائه می کند که پیچش فرکانسی را به صورت درون یابی خطی انرژی های فیلتر بانک های بهم پیوسته ی میل (Mel) مدل می کند. هدف این روش کاهش اعوجاجات در تخمین انرژی فیلتر بانک میل است که به علت ترکیب هارمونیک فواصل زمانی گفتار واکدار<sup>۱</sup> و نمونه برداری DFT (تبدیل فوریه گسسته)، هنگامی که فرکانس مرکزی فیلتر باندگذر شیفت پیدا کرده، ایجاد شده است.

در تمامی روش های بالا لازم است تا ضریب پیچش فرکانسی برای تمامی گویندگان تخمین زده شود. برای این کار رویکردهای مختلفی پیشنهاد شده است. به عنوان مثال در (ایدی، ۱۹۹۶) پیشنهاد شده است که با استفاده از فورمنت های موجود در سیگنال گفتار، تخمینی از طول مجرای صوتی به دست آید و با استفاده از این تخمین، ضریب پیچش فرکانسی محاسبه شود. مشکل اصلی این دسته از روش ها که مبتنی بر تخمین طول مجرای صوتی هستند، خطای ذاتی موجود در همین تخمین است. چون این تخمین خود وابسته به آشکارسازی فورمنت ها بوده و در نتیجه نیازمند مشخص کردن نواحی صدادر گفتار است (وانگ، ۲۰۰۷).

رویکرد رایج دیگر، جستجوی شبکه ای<sup>۲</sup> (لی، ۱۹۹۸؛ وگمان، ۱۹۹۶) است. در این روش، از میان مجموعه ای از ضرایب پیچش که به طور معمول با فواصل یکسان در یک بازه مشخص می شوند، ضریبی که بیشترین دقت بازشناسی را به دست می دهد انتخاب می شود. این کار هزینه محاسباتی بسیار بالایی دارد؛ چون برای هر ضریب باید یک بار عملیات بازشناسی انجام شود و همچنین محدودیت پیاده سازی از لحاظ نیاز به برچسب دقیق فایل صوتی مورد پردازش دارد. از این رو به طور معمول به جای بازشناسی، از جستجوی شبکه ای مبتنی بر بیشینه سازی درست نمایی استفاده می شود که در این روش به جای بیشینه سازی دقت بازشناسی، درست نمایی بردارهای ویژگی که پیچش بر روی آنها اعمال شده به مدل بازشناسی، بیشینه می شود (لی،

<sup>3</sup> Expectation-Maximization

<sup>4</sup> Gaussian Mixture Model

<sup>1</sup> Voiced Speech

<sup>2</sup> grid search



(شکل-۱): به‌کارگیری روش VTLN در روش استخراج ویژگی PLPC

هر سامانهٔ بازشناسی دارای دو مرحله آموزش و آزمون (رمزگشایی) است. در مرحلهٔ آموزش مدل صوتی، ابتدا بایستی ضرایب پیش‌چشم تمام گوینده‌های آموزشی را به‌دست آورد؛ سپس در مرحلهٔ استخراج ویژگی، بعد از محاسبه طیف، این ضرایب را جهت پیش‌چشم طیف اعمال کرد تا بردار ویژگی‌های جدید محاسبه شوند و سپس با استفاده از این بردارهای ویژگی، مدل صوتی نهایی تعلیم یابد. در مرحلهٔ به‌کارگیری سامانهٔ (آزمون) هم ابتدا باید برای سیگنال ورودی ضریب پیش‌چشم مناسب را پیدا کرد و بعد از اعمال آن، بایستی با بهره‌گیری از مدل هنجار شده صوتی، فرایند بازشناسی انجام گیرد.

### ۳- تعیین ضریب پیش‌چشم

همان‌طور که در بخش ۲ ذکر شد، برای پیاده‌سازی روش VTLN ابتدا باید ضریب پیش‌چشم را برای هر گوینده از دادگان آموزش محاسبه کرد و سپس ویژگی هنجار شده به طول مسیر صوتی استخراج شود.

تعیین مقدار مناسب ضریب پیش‌چشم با یک الگوریتم جستجو انجام می‌شود. می‌دانیم که مقدار ضریب پیش‌چشم در یک محدوده کوچک اطراف مقدار یک قرار دارد (ویلینگ، ۲۰۰۲)؛ لذا با تغییر مقدار ضریب پیش‌چشم در این محدوده و با گام تغییرات به اندازهٔ کافی کوچک (به‌طور معمول برابر با

سریع ضرایب پیش‌چشم گوینده پیشنهاد شده است که مبتنی بر امتیازهای صوتی به‌دست‌آمده از سامانهٔ تشخیص جنسیت گوینده است و نه تنها دقت بازشناسی را نسبت به روش جستجوی شبکه‌ای کاهش نمی‌دهد، بلکه منجر به بهبود آن نیز می‌شود.

ساختار مقاله به شرح زیر است: در بخش ۲ روش هنجارسازی طول مجرای صوتی معرفی می‌شود. در بخش ۳ نحوه یافتن ضرایب پیش‌چشم و آموزش مدل هنجارسازی شده شرح داده می‌شود. در بخش ۴ سامانه تشخیص جنسیت گوینده مورد استفاده بیان خواهد شد. در بخش ۵ روش پیشنهادی و نحوهٔ به‌دست‌آوردن ضریب پیش‌چشم با استفاده از سامانه تشخیص جنسیت توضیح داده می‌شود. دادگان مورد استفاده در بخش ۶ معرفی شده و در بخش ۷ سامانه پایه، آزمایش‌های پیاده‌سازی شده و بحث و بررسی نتایج به‌دست‌آمده گنجانده می‌شود و در نهایت در بخش ۸ به نتیجه‌گیری مقاله خواهیم پرداخت.

## ۲- معرفی روش هنجارسازی طول مجرای صوتی (VTLN)

روش VTLN می‌تواند با پیش‌چشم<sup>۱</sup> محور فرکانسی (FW) در بخش تحلیل فیلتر بانک در دو روش متداول استخراج ویژگی گفتاری MFCC یا PLPC پیاده‌سازی شود. در شکل (۱) روند نامی روش VTLN در روش استخراج ویژگی PLPC نشان داده شده است.

پیش‌چشم محور فرکانسی می‌تواند با ضرایب مختلفی انجام شود که این ضرایب بیان‌گر تغییرات موجود در طول لولهٔ صوتی گوینده‌ها می‌باشند. در واقع هر گوینده دارای ضریب پیش‌چشم خاص خود می‌باشد که مقدار آن متناسب با عکس طول لولهٔ صوتی آن گوینده است. در نتیجه برای پیاده‌سازی روش VTLN باید طیف گفتار هر گوینده با ضریب پیش‌چشم خاص آن گوینده مقیاس‌دهی شود. برای پیاده‌سازی این روش، به‌طور متداول سه مرحله متوالی زیر انجام می‌گیرد:

- ۱- پیدا کردن ضریب پیش‌چشم بهینه برای سیگنال گفتار هر گوینده.
- ۲- پیش‌چشم فرکانسی طیف گفتار و هنجارسازی آن با ضریب پیش‌چشم بهینه.
- ۳- استفاده از طیف هنجارسازی شده در بخش استخراج ویژگی.

<sup>1</sup> Warping

**تابع تکه‌ای-خطی:** این رابطه خطی بوده و به صورت زیر مشخص می‌شود:

$$f_{\alpha}(w) = \begin{cases} \alpha w & w \leq \omega_0 \\ \alpha \omega_0 + \frac{\pi - \alpha \omega_0}{\pi - \omega_0} (w - \omega_0) & w \geq \omega_0 \end{cases} \quad (4)$$

این تابع دو پارامتر  $\alpha$  و  $\omega_0$  دارد که  $\omega_0$  نقطه شکستگی در شکل (۲.الف) می‌باشد و به طور معمول توسط رابطه زیر تعیین می‌شود:

$$\omega_0 = \begin{cases} \frac{7}{8}\pi & \alpha \leq 1 \\ \frac{7}{8\alpha}\pi & \alpha \geq 1 \end{cases} \quad (5)$$

**تابع دوسویگی:** این رابطه غیر خطی بوده و فقط شامل یک پارامتر  $\beta$  می‌باشد و در شکل (۲.ب) نشان داده شده است.

$$f_{\beta}(w) = \omega + 2 \cdot \tan^{-1} \left\{ \frac{(1 - \beta) \sin \omega}{1 - (1 - \beta) \cos \omega} \right\} \quad (6)$$

در قسمت‌هایی از منحنی که مشتق نخست تابع بالا دارای مقدار عددی کمتر از یک باشد، منجر به فشردن محور فرکانسی (بیش تر در صداهای زیر مانند گفتار زنان) می‌شود و برای مقادیر بزرگتر از یک، معادل با بسط دادن (باز کردن) محور فرکانسی (بیش تر در صداهای بم مانند گفتار مردان) خواهد بود.

در صورت استفاده از توابع  $f_{\alpha}$  یا  $f_{\beta}$  برای پیچش فرکانسی، منظور از تخمین ضریب پیچش، تخمین پارامترهای این توابع ( $\beta$  و  $\alpha$ ) خواهد بود. با توجه به اینکه استفاده از تابع تکه‌ای-خطی ( $f_{\alpha}$ ) رایج تر است، در این مقاله نیز از این تابع استفاده می‌شود و در ادامه مقاله منظور از ضریب پیچش، پارامتر  $\alpha$  است.

بعد از به دست آوردن ضریب پیچش برای هر گوینده، این ضریب به طیف گفتار جهت پیچش اعمال شده و ویژگی هنجار شده گفتار استخراج می‌شود؛ سپس با استفاده از این ویژگی‌های هنجار شده، مدل صوتی هنجار شده ساخته می‌شود.

ضریب پیچش گفتار در مرحله آزمون نیز همانند مراحل آموزش محاسبه می‌شود؛ سپس با استفاده از این ضرایب، ویژگی‌های هنجار شده به دست می‌آیند و در نهایت بازنشانی گفتار بر روی ویژگی‌های هنجار شده و با استفاده

(۰/۰۱) در مرحله استخراج ویژگی و سپس محاسبه امتیاز لگاریتم درست‌نمایی<sup>۱</sup> (LL) در مرحله بازنشانی، ضریب پیچش بهینه تعیین می‌شود. کوهن در (کوهن، ۱۹۹۴) نشان داد که ضریب پیچش مناسب منجر به بیشینه کردن امتیاز LL می‌شود.

نحوه اعمال ضریب پیچش در روش استخراج ویژگی PLPC بدین شرح است: فرض کنیم  $X(w)$  طیف نخستین به دست آمده از یک قاب گفتاری باشد. از آنجا که در روش استخراج ویژگی PLPC از فیلترهای دوزنقه ای بارک  $T(w)$  در مرحله فیلتر بانک استفاده می‌شود، خروجی فیلتر  $n$ ام بدون پیچش به صورت رابطه زیر درمی‌آید:

$$O(n) = \sum_{\omega=l_n}^{\omega=h_n} T_n(\omega) \cdot X(\omega) \quad 0 \leq n \leq N-1$$

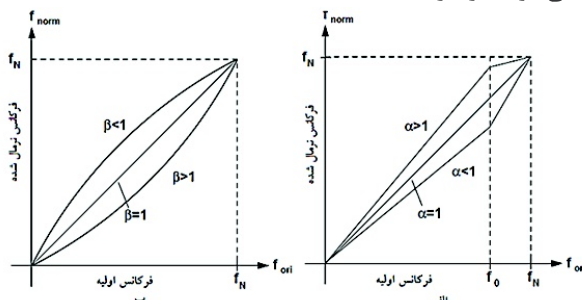
که در آن  $N$  تعداد فیلترها،  $h_n$  و  $l_n$  به ترتیب فرکانس قطع پایین و فرکانس قطع بالای فیلتر  $n$ ام است. اگر تابع  $f$  وظیفه پیچش محور فرکانسی را داشته باشد، در این صورت برای طیف انتقال یافته داریم:

$$Y(\omega) = X(f(\omega))$$

این مسئله (پیچش محور فرکانسی) در شکل ۲ نشان داده شده است که در آن فرکانس نخستین  $f_{ori}$  بعد از اعمال پیچش به فرکانس  $f_{norm}$  تبدیل شده است. در نهایت خروجی فیلتر بانک پیچش خورده به صورت زیر محاسبه می‌شود:

$$O_f(n) = \sum_{\omega=l_n}^{\omega=h_n} T_n(\omega) \cdot X(f(\omega)) \quad 0 \leq n \leq N-1 \quad (3)$$

به طور معمول برای تابع  $f$  دو نوع رابطه متداول استفاده می‌شود (مولانو، ۲۰۰۰):



(شکل-۲): هنجار سازی طول مسیر صوتی و پیچش طیف سیگنال گفتاری. الف) تابع تکه‌ای-خطی. ب) تابع دوسویگی (مولانو، ۲۰۰۰).

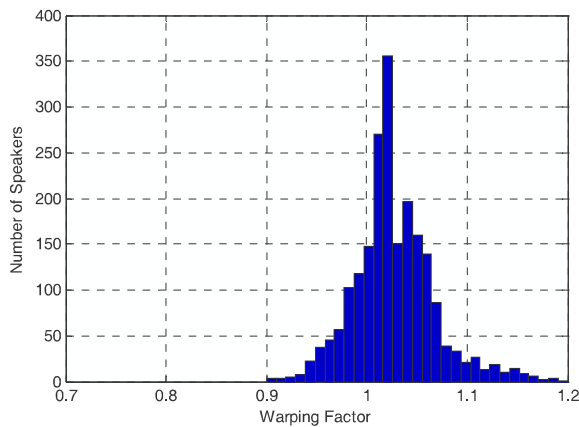
<sup>2</sup> Compressed spectrum

<sup>3</sup> Expanded spectrum

<sup>1</sup> Log-Likelihood



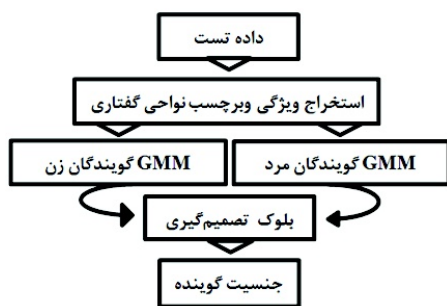
گوینده تخمین زده شود. در ادامه مقاله سامانه مورد استفاده برای تشخیص جنسیت گوینده معرفی می‌شود.



(شکل - ۴): نمودار هیستوگرام توزیع ضرایب پیچش بهینه برای گویندگان مرد

#### ۴- سامانه تشخیص جنسیت گوینده

در این بخش روش و اجزای سامانه طراحی شده برای تشخیص جنسیت گوینده معرفی می‌شود. سامانه تشخیص جنسیت گوینده مبتنی بر روش بیز<sup>۱</sup> از سه بخش اصلی استخراج ویژگی، محاسبه امتیاز و تصمیم‌گیری تشکیل می‌شود. روش کار بدین صورت است که با استفاده از بردارهای ویژگی به دست آمده از بخش‌های گفتاری نمونه‌های متنوعی از گویندگان زن و مرد، برای هر دو جنس به‌طور مجزا مدل مخلوط گوسی (GMM) تعلیم داده می‌شود و سپس با محاسبه احتمال تعلق داده، آزمون به دو مدل، مدل برنده تعیین و جنسیت گوینده مشخص می‌شود. در شکل (۵) شماتیک کلی این سامانه نمایش داده شده است.



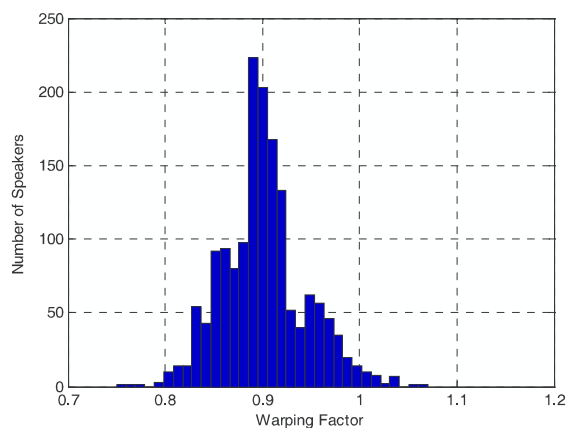
(شکل - ۵): بلوک دیاگرام سیستم تشخیص جنسیت گوینده

در زیربخش‌های بعدی ابتدا نحوه استخراج ویژگی و سپس روش تعلیم تمایزی GMM-MMI برای آموزش مدل‌های جنسیتی توضیح داده؛ سپس دادگان مورد استفاده

از مدل صوتی آموزش یافته بر روی دادگان گفتاری تعلیم هنجار شده، انجام می‌گیرد.

از دید عملیاتی، مهم‌ترین و پرهزینه‌ترین بخش برای پیاده‌سازی VTLN، مرحله نخست، یعنی یافتن ضریب پیچش بهینه برای هر گوینده است (جستجوی بهترین ضریب از میان ۵۱ ضریب که به معنی ۵۱ بار محاسبه امتیاز درست‌نمایی است)؛ لذا روش VTLN مبتنی بر جستجو به روشی با هزینه محاسباتی بالا بدل خواهد شد. بنابراین در این مقاله روش پیشنهاد می‌شود که هزینه محاسباتی و زمان لازم برای تعیین ضریب پیچش مناسب را به شدت کاهش دهد.

تجربه نشان داده است که ضرایب پیچش، ارتباط مستقیمی با جنسیت گوینده دارند و ضرایب پیچش برای گویندگان زن به‌طور معمول کمتر از ۱ و برای گویندگان مرد به‌طور عمومی بیشتر از یک است. در شکل‌های ۳ و ۴ این ضرایب برای گویندگان زن و مرد مربوط به دادگان تعلیم با استفاده از روش جستجو که در این مقاله به‌عنوان روش بهینه در نظر گرفته شده است، به دست آمده و توزیع آنها رسم شده است. محور عمودی این توزیع‌ها، تعداد گوینده‌هایی را نشان می‌دهد که ضریب پیچش آنها برابر عدد مشخص شده در محور افقی است. (مشخصات دادگان به کار رفته و شرایط آزمایش‌ها در بخش ۶ آمده است).



(شکل - ۳): نمودار هیستوگرام توزیع ضرایب پیچش بهینه برای گویندگان زن

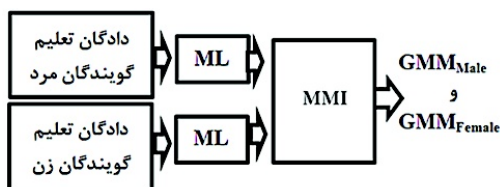
در این مقاله با استفاده از این واقعیت که نحوه توزیع ضرایب پیچش ارتباط مستقیمی با جنسیت گوینده دارد و همچنین کم‌هزینه بودن تعیین خودکار جنسیت گوینده، برای کاهش هزینه محاسباتی تعیین ضرایب پیچش، روش جدیدی پیشنهاد شده است که با استفاده از امتیاز به دست آمده از سامانه تشخیص جنسیت گوینده، ضریب پیچش هر

<sup>1</sup> Bayes

$$F_{MMI}(\lambda) = \sum_{r=1}^R \log \frac{p_{\lambda}(x_r | G_r)p(G_r)}{\sum_{\forall G} p_{\lambda}(x_r | G)p(G)} \quad (8)$$

در مخرج رابطه فوق، عبارت  $(\sum_{\forall G} p(x_r | G))$ ، احتمال گفتار  $x_r$  با در نظر گرفتن مدل رقیب (جنس مخالف) است. بیش تر احتمال پیشین کلیه طبقاتها یکسان در نظر گرفته می شود. بنابراین جمله های  $p(G_r)$  و  $p(G)$  از رابطه بالا حذف می شوند.

در این مقاله ابتدا مدل مربوط به دو جنس با استفاده از الگوریتم ML تعلیم یافته و سپس با استفاده از روش MMI پارامترهای مدل ها مجدداً تعلیم داده می شود تا تمایز مدل دو جنس افزایش یابد (شکل ۶).



(شکل - ۶): نحوه تعلیم تمایزی مدل وابسته به دو جنس مرد ( $GMM_{Male}$ ) و زن ( $GMM_{Female}$ ).

#### ۴-۳- مرحله ارزیابی در تشخیص جنسیت

در روش GMM-MMI، برای تصمیم گیری در مورد داده آزمون، امتیاز تشخیص جنسیت زیر محاسبه می شود:

$$GD_{test} = \log p(x | GMM_{Male}) - \log p(x | GMM_{Female})$$

if  $GD_{test} > \theta \Rightarrow \text{Male}$   
if  $GD_{test} < \theta \Rightarrow \text{Female}$

در این رابطه  $x$  ویژگی های گفتاری استخراج شده از داده آزمون است که شباهت آن به مدل گویندگان مرد ( $GMM_{Male}$ ) و مدل گویندگان زن ( $GMM_{Female}$ ) محاسبه می شود. امتیاز نهایی ( $GD_{test}$ ) با یک مقدار آستانه از پیش تعیین شده ( $\theta$ ) مقایسه و سپس در مورد جنسیت داده آزمون تصمیم گیری می شود.

#### ۵- معرفی روش پیشنهادی: بررسی

#### ارتباط ضرایب پیچش بهینه با امتیاز مدلسازی جنسیت

همان طور که در قبل اشاره شد، در روش های متداول پیاده سازی الگوریتم VTLN تعیین مقدار بهینه ضریب

برای تعلیم و آزمون و دقت سیستم تشخیص جنسیت گوینده بیان می شود.

#### ۴-۱- استخراج ویژگی

استخراج ویژگی: برای سامانه تشخیص جنسیت گوینده از ویژگی های کپستروم متداول در بازشناسی گفتار استفاده شده است. برای بهبود کیفیت ویژگی های استخراج شده، فیلتر RASTA به بردارهای ویژگی استخراج شده اعمال شده است. این فیلتر با حذف فرکانس های نزدیک به صفر (در حوزه فرکانس مدولاسیون)، موجب کاهش برخی فرکانس های کوچک مخرب (نوفه های ایستان و اعوجاجات ثابت کانال) از سیگنال می شود. از طرف دیگر این فیلتر فرکانس های بالا و غیر مفید در حوزه فرکانس مدولاسیون را نیز حذف می کند.

#### ۴-۲- روش مدل سازی جنسیتی مبتنی بر GMM-MMI

تشخیص جنسیت گوینده یک مسئله طبقه بندی الگوی دو طبقه (دو جنس زن و مرد) است که به طور معمول با روش بیز پیاده سازی می شود. یکی از روش های معمول برای مدل سازی توزیع دادگان هر طبقه، استفاده از مدل مخلوط گوسی تعلیم یافته با روش ML است. در روش تعلیم استاندارد ML، مجموع لگاریتم احتمال کل دادگان تعلیم، به عنوان تابع هدف در نظر گرفته می شود و هدف بیشینه کردن تابع ذیل است (رینالد، ۲۰۰۰):

$$F_{ML}(\lambda) = \sum_{r=1}^R \log p(x_r | G_r, \lambda) \quad (9)$$

در این رابطه  $\lambda$  پارامترهای مدل،  $x_r$  نشان گر  $r$  امین قاب گفتار تعلیم،  $R$  تعداد قاب های گفتار تعلیم و  $G_r$  برچسب طبقه داده تعلیم  $r$  ام (جنسیت گوینده) است. برای بیشینه کردن تابع هدف، پارامترهای GMM به صورت تکراری و با استفاده از روابط تخمین الگوریتم EM محاسبه می شود.

پس از آموزش مدل با روش ML، با استفاده از تعلیم تمایزی، دقت بازشناسی بهبود داده می شود. یکی از الگوریتم های تعلیم تمایزی مشهور، الگوریتم تعلیم MMI<sup>۱</sup> است که در آن تابع هدف زیر بیشینه می شود (بورگت، ۲۰۰۶):

<sup>1</sup> Maximum Mutual Information

در ادامه با استفاده از تابع به دست آمده، مقدار ضریب پیچش برای هر گوینده، از روی امتیاز تشخیص جنسیت تخصیص یافته به آن گوینده محاسبه شده و این ضریب به طیف گفتار جهت پیچش اعمال شده و ویژگی هنجار شده گفتار استخراج می شود.

## ۶- معرفی دادگان

دادگان گفتاری مورد استفاده برای تعلیم مدل صوتی، شامل سه مجموعه از گفتار فارسی است که برای کانال تلفنی آماده شده است. مجموعه دادگان فارسی دات تلفنی کوچک (بی جن خان، ۲۰۰۳) که شامل جملات بیان شده از ۶۴ نفر گوینده زن و مرد است و در مجموع حدود ۸ ساعت گفتار را در بر می گیرد. دادگان فارسی دات بزرگ (۷۴ ساعت) (شیخ زادگان، ۲۰۰۶) که حاوی یکصد گویشور است و به طور متوسط هر گویشور ۲۰ تا ۲۵ فایل ۱/۵ تا ۲ دقیقه ای از متن های رسمی که بیش تر روزنامه ای هستند، خوانده است. همچنین از یک مجموعه ۱۰۲ ساعته دادگان فارسی حاوی دوپست گوینده استفاده شده است که در آن هر گوینده شش فایل در حدود پنج دقیقه ای را از روی متون رسمی و نیمه محاوره ای خوانده است. این دادگان به منظور گسترش و تکمیل دادگان فارسی دات بزرگ و با همان شرایط تهیه شده است. با توجه به اینکه آزمایش های این مقاله در شرایط تلفنی گزارش شده اند دو مجموعه داده آخر از خطوط تلفن عبور داده شده اند تا شرایط کانال آن را داشته باشند. دادگان گفتاری مورد استفاده برای آزمون نیز شامل ۹۴ دقیقه از گفتار نیمه محاوره ای تلفنی فارسی است که توسط ۱۰ گوینده زن و ۱۰ گوینده مرد (متفاوت از گویندگان آموزشی) از روی متون نیمه محاوره ای خوانده شده است.

مدل زبانی به کاررفته برای بازشناسی دادگان آزمون نیمه محاوره ای، ۳-گرامی<sup>۲</sup> است که حاوی حدود ۱۱/۵ میلیون ۳-گرام است که از متون رسمی- محاوره ای با حجم در حدود یکصد میلیون کلمه ای (برگرفته از پیکره متنی فارسی) و چندین متن محاوره ای در حدود ۱۴ میلیون کلمه- ای ساخته شده است. مجموعه واژگان مورد استفاده نیز شامل حدود ۵۳ هزار کلمه با میانگین تنوع تلفظی ۲/۴ تلفظ برای هر کلمه است. تنوعات تلفظی هر کلمه در مجموعه واژگان براساس ۲۹ واج متداول فارسی تهیه شده است (شکفته، ۲۰۱۳).

پیچش با یک الگوریتم جستجو انجام می شود که از لحاظ محاسباتی کاری هزینه بر است. روش پیشنهادی این مقاله برای کاهش محاسبات بدین صورت است که با استفاده از امتیاز به دست آمده از سامانه تشخیص جنسیت گوینده، پارامتر بهینه پیچش تخمین زده شود. برای این کار به دنبال پیدا کردن تابع مناسبی هستیم که با استفاده از امتیاز تشخیص جنسیت به دست آمده از رابطه (۹)، تخمین مناسبی از مقدار ضریب پیچش ارائه کند. برای این منظور از معیار کمینه مربعات خطا<sup>۱</sup> برای تخمین تابع بر روی تمامی دادگان آموزشی استفاده می کنیم. فرض کنیم تابع انتخابی به صورت چند جمله ای درجه یک، به صورت زیر در نظر گرفته شود:

$$WF(GD) = a_1 * GD + a_0 \quad (10)$$

که در آن  $WF$  ضریب پیچش بهینه،  $GD$  امتیاز تشخیص جنسیت،  $a_0$  و  $a_1$  ضرایب رگرسیون خطی می باشند.

با فرض اینکه  $R$  نقطه به صورت  $(gd_r, wf_r)$  داشته باشیم، تقریب حداقل مربعات خطا به صورت زیر محاسبه می شود:

$$S = \sum_{r=1}^R (wf_r - WF(gd_r))^2 = \sum_{r=1}^R (wf_r - (a_1 * gd_r + a_0))^2 \quad (11)$$

در این رابطه  $R$  تعداد دادگان آموزشی،  $wf_r$  ضریب پیچش بهینه برای داده آموزشی شماره  $r$ ،  $gd_r$  امتیاز تشخیص جنسیت برای همان داده آموزشی و  $S$  میانگین مربعات خطا است. در این روابط متغیرهای مستقل با حروف کوچک و متغیرهای وابسته با حروف بزرگ مشخص شده اند. مقدار ضرایب  $a_0$  و  $a_1$  باید طوری محاسبه شود که  $S$  کمینه شود. در نتیجه ضرایب از حل معادله زیر به دست می آید:

$$\frac{\partial S}{\partial a_0} = 0, \frac{\partial S}{\partial a_1} = 0 \quad (12)$$

در نتیجه بعد از ساده سازی، ضرایب به صورت زیر به دست می آیند:

$$a_0 = \frac{\sum_{r=1}^R wf_r \sum_{r=1}^R gd_r^2 - \sum_{r=1}^R gd_r \sum_{r=1}^R gd_r wf_r}{R \sum_{r=1}^R gd_r^2 - (\sum_{r=1}^R gd_r)^2} \quad (13)$$

$$a_1 = \frac{R \sum_{r=1}^R gd_r wf_r - \sum_{r=1}^R gd_r \sum_{r=1}^R wf_r}{R \sum_{r=1}^R gd_r^2 - (\sum_{r=1}^R gd_r)^2} \quad (14)$$

<sup>2</sup> Trigram

<sup>1</sup> Least Square Error



مجموعه مدل صوتی تک‌آوایی<sup>۲</sup> بر مبنای مدل مخفی مارکوف تعلیم داده می‌شود؛ سپس در مرحله استخراج ویژگی، با تغییر مقدار ضریب پیچش در محدوده (۰/۷۰-۰/۸۰) و با گام تغییرات ۰/۰۱، ۵۱ سری ویژگی‌های مختلف به دست می‌آید و برای هر کدام از این سری ویژگی‌ها، با استفاده از مدل صوتی تک‌آوایی، بازنمایی در سطح واج انجام می‌گیرد و امتیاز لگاریتم درست‌نمایی محاسبه می‌شود. در نهایت ضریب پیچشی که امتیاز لگاریتم درست‌نمایی را بیشینه کند، به عنوان ضریب پیچش بهینه انتخاب می‌شود. این عملیات بر روی کل مجموعه آموزش انجام می‌شود.

در ادامه با استفاده از ضریب پیچش بهینه، ویژگی‌های هنجار شده کل گفتار هر گوینده استخراج و با استفاده از ویژگی‌های هنجار شده دادگان آموزشی، مدل صوتی نهایی تعلیم داده می‌شود. در نهایت نیز ویژگی‌های هنجار شده دادگان آزمون با استفاده از مدل صوتی ساخته شده در سطح کلمه بازنمایی می‌شود.

## ۷-۲- روش مبتنی بر جدول گویندگان (میمورا، ۲۰۱۱)

در این روش برای تعدادی گوینده مرجع، ضریب پیچش با استفاده از روش مبتنی بر جستجو محاسبه می‌شود؛ سپس برای داده‌های آزمایش با استفاده از روش  $\Delta BIC$  نزدیک‌ترین گوینده از مجموعه بالا مشخص شده و از ضریب پیچش مربوط به آن گوینده برای داده آزمایش استفاده می‌شود. برای پیاده‌سازی این روش از تمامی گویندگان موجود در دادگان آموزش (۳۶۴ گوینده) به عنوان گویندگان مرجع استفاده کرده‌ایم؛ سپس به هر یک از ۲۰ گوینده دادگان آزمون، ضریب پیچش نزدیک‌ترین گوینده از مجموعه مرجع اختصاص داده شده است.

## ۷-۳- روش پیشنهادی-مبتنی بر امتیاز جنسیت

با استفاده از رابطه ۱۰ در بخش ۵، تخمین ضرایب پیچش جدید از روی امتیاز تشخیص جنسیت تعیین می‌شود. ویژگی‌های هنجار شده داده‌های آموزش با استفاده از این ضرایب پیچش جدید استخراج می‌شوند و مدل صوتی با استفاده از آنها ساخته می‌شود. در نهایت ویژگی‌های هنجار شده دادگان آزمون با استفاده از این مدل، بازنمایی می‌شوند. این روش در دو حالت پیاده‌سازی شده است: در حالت نخست برای کل گویندگان از یک مدل رگرسیون (ضرایب  $a_0$  و  $a_1$  در رابطه ۱۰)

برای تعلیم سامانه تعیین جنسیت گوینده از دادگان فارس‌دات بزرگ (تلفنی شده) استفاده شده است. دادگان مورد استفاده برای آزمون سامانه نیز دادگان فارس‌دات کوچک تلفنی بوده است. دقت سامانه تشخیص جنسیت گوینده بر روی این دادگان ۹۵/۹۴ درصد است که از تقسیم تعداد فایل‌های درست به کل فایل‌ها به دست آمده است.

## ۷- معرفی سامانه پایه، آزمایش‌های انجام شده و بحث و بررسی نتایج

برای پیاده‌سازی روش VTLN و تعلیم مدل صوتی و بازنمایی دادگان آزمون در این مقاله از سامانه بازنمایی گفتار پیوسته با مدل‌های صوتی مبتنی بر مدل مخفی مارکوف (HMM) به کمک جعبه ابزار HTK استفاده شده است (<http://htk.eng.cam.ac.uk>). این سامانه، شامل مجموعه‌ای از مدل‌های HMM سه‌آوایی<sup>۱</sup> چپ به راست برای هر یک از ۲۹ واج فارسی است، که به طور متوسط برای هر واج تعداد چهار حالت (State) در نظر گرفته می‌شود. توزیع احتمالاتی ویژگی‌ها در هر حالت نیز با مدل مخلوط گوسی (GMM) با تعداد ۳۲ گوسی در نظر گرفته شده است. ویژگی‌های استخراج شده برای هر قاب گفتار تلفنی، شامل ۱۲ ضریب PLPC به همراه ضریب کپسترال  $C_0$  است و از مشتقات مرتبه نخست تا سوم ضرایب کپستروم به دست آمده برای مدل‌سازی دینامیک ویژگی‌های گفتاری استفاده شده است. همچنین در جهت مقاوم سازی بردار ویژگی به دست آمده نسبت به نوفه‌های جمع‌شونده (مانند صدای محیط) و اثر انتقال کانال تلفنی از روش پس پردازش MVA (۱۳۸۹). نتایج بازنمایی بر حسب درصد دقت (Accuracy) بازنمایی کلمه گزارش شده است.

برای پیاده‌سازی روش VTLN، برای هر گفتار گوینده، ضریب پیچش بهینه با استفاده از روش مبتنی بر جستجو، روش مبتنی بر جدول گویندگان و روش پیشنهادی، به شرح زیر محاسبه شد:

## ۷-۱- روش مبتنی بر جستجو ((باباعلی، (۱۳۸۶)):

ابتدا با استفاده از ویژگی PLPC با بُعد ۵۲ و با ضریب پیچش یک (معادل با عدم اعمال ضریب پیچش)، یک

<sup>2</sup> monophone

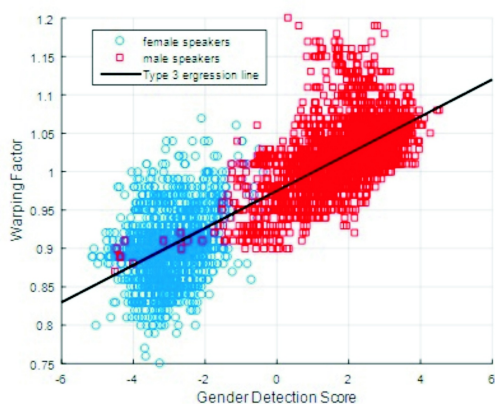
<sup>1</sup> Triphone

یک ماشین مجازی با ۱۲ هسته ۲/۵ گیگاهرتزی و ۱۲ گیگابایت حافظه RAM انجام شده‌اند. این ماشین مجازی بر روی یک رایانه سرور با پردازنده Xenon ساخته شده است. آموزش مدل‌های صوتی بر روی این سامانه با استفاده از قابلیت پردازش موازی جعبه ابزار HTK، در حدود پانزده ساعت زمان صرف کرده است.

(جدول-۲): زمان صرف شده برای پردازش یک دقیقه از گفتار بر حسب ثانیه به همراه درصد سهم هر بخش.

نوع	زمان استخراج و ویژگی		زمان تخمین ضریب		زمان بازشناسی	
	درصد	زمان	درصد	زمان	درصد	زمان
نوع ۱	۰/۱۷	%۰/۰۷	۱۱۵/۰۶	%۴۸/۳۹	۱۲۲/۵۵	%۵۱/۵۴
نوع ۲	۰/۱۷	%۰/۱۴	۰/۸۹	%۰/۷۵	۱۱۸/۰۹	%۹۹/۱۱
نوع ۳ و ۴	۰/۱۷	%۰/۱۵	۰/۶۸	%۰/۵۹	۱۱۴/۸۹	%۹۹/۲۶

در شکل (۷) توزیع ضریب پیش گفتارهای آموزشی بر حسب امتیاز تشخیص جنسیت آنها به همراه مدل خطی به دست آمده (نوع ۳)، نشان داده شده است. همان‌طور که مشاهده می‌شود برای گویندگان زن، ضریب پیش آنها مقداری کمتر از یک و امتیاز جنسیت آنها منفی است. همچنین برای اکثریت گویندگان مرد، ضریب پیش آنها مقداری بیشتر از یک و امتیاز جنسیت آنها مثبت است.



(شکل-۷): توزیع ضریب پیش بهینه مجموعه متنوعی از دادگان آموزشی بر حسب امتیاز تشخیص جنسیت آنها. مربع‌های قرمز و دایره‌های آبی رنگ به ترتیب نشان‌دهنده گویندگان مرد و زن هستند و خط سیاه نشان‌دهنده منحنی به دست آمده از رگرسیون در حالت نوع سه است.

و در حالت دوم از دو مدل رگرسیون جداگانه برای گویندگان مرد و زن استفاده شده است.

## ۷-۴- پیاده‌سازی آزمایش‌ها

برای پیاده‌سازی آزمایش‌ها چهار نوع ضریب پیش متفاوت محاسبه خواهد شد که در جدول زیر آورده شده است:

(جدول-۱): معرفی انواع ضریب پیش پیاده‌سازی شده در آزمایش‌ها

حالت پایه	بدون ضریب پیش (WF=1)
نوع ۱	محاسبه ضریب پیش با استفاده از امتیاز لگاریتم درست نمایی از روی کل واج‌های بازشناسی شده (روش پایه VTLN مطابق با مرجع (باباعلی و همکاران ۱۳۸۶))
نوع ۲	محاسبه ضریب پیش با روش مبتنی بر جدول گویندگان (میمورا، ۲۰۱۱)
نوع ۳	محاسبه ضریب پیش با روش تخمین رابطه ۱۰ (روش پیشنهادی مقاله)
نوع ۴	محاسبه ضریب پیش با روش تخمین رابطه ۱۰ به صورت جداگانه برای گویندگان مرد و زن (روش پیشنهادی مقاله)

جدول (۲)، زمان پردازش یک دقیقه گفتار را برای هر یک از روش‌های جدول (۱) و به تفکیک بخش‌های مختلف آنها بر حسب ثانیه نشان می‌دهد. همچنین در این جدول سهم هر بخش از پردازش کل در مرحله آزمون نیز به صورت درصد بیان شده است.

همان‌طور که در قبل نیز اشاره شد، به دست آوردن ضریب پیش بهینه با استفاده از روش مبتنی بر الگوریتم جستجو (نوع ۱) زمان‌بر است؛ به طوری که برای محاسبه ضریب پیش بهینه، برای یک فایل گفتار یک دقیقه‌ای زمانی در حدود ۱۱۵ ثانیه لازم است که معادل ۴۸ درصد از زمان کل پردازش است؛ در حالی که برای به دست آوردن ضریب پیش هر دقیقه گفتار با استفاده از روش پیشنهادی (نوع ۳ و ۴)، زمانی در حدود ۰/۶۸ ثانیه مورد نیاز خواهد بود (کمتر از ۱ درصد از پردازش کل). روش مبتنی بر جدول گویندگان (نوع ۲) نیز از نظر هزینه محاسبات، عملکرد مناسبی داشته و برای هر دقیقه گفتار حدود ۰/۸۹ ثانیه زمان صرف کرده است. برنامه‌های مورد نیاز برای به دست آوردن ضریب پیش در محیط MATLAB پیاده‌سازی شدند و آزمایش‌ها بر روی

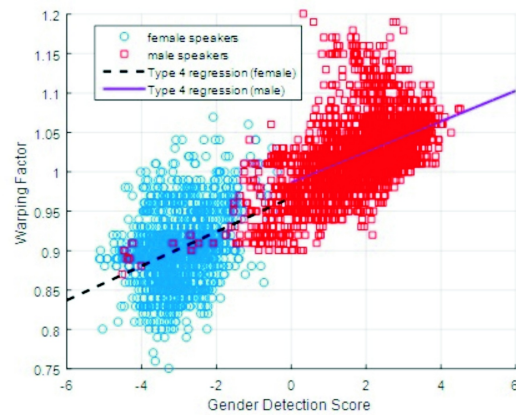
روش مبتنی بر جدول گویندگان (نوع ۲)، بسیار نزدیک به روش مبتنی بر جستجو (نوع ۱) بوده است. این موضوع بدین صورت قابل توجیه است که در این روش برای گوینده آزمون به طور دقیق از ضریب پیچش گوینده‌ای استفاده می‌شود که ضریب پیچش آن توسط روش نوع یک به دست آمده است. با این حال، سرعت این روش نیز بسیار بالا بوده و در حدود ۱۲۹ مرتبه سریع‌تر از روش مبتنی بر جستجو است.

(جدول - ۳): نتایج درصد دقت بازشناسی کلمه با انواع متفاوت تعیین ضریب پیچش در دادگان تعلیم و آزمون و به تفکیک جنسیت گویندگان

نوع ضریب پیچش دادگان آموزش و آزمون	درصد دقت بازشناسی کلمه		
	کل	مرد	زن
حالت پایه (WF=1)	۵۴/۵۵	۵۸/۶۷	۵۰/۳۲
نوع ۱	۵۵/۹۲	۵۹/۴۰	۵۲/۳۳
نوع ۲	۵۵/۹۴	۵۹/۳۸	۵۲/۴۰
نوع ۳	۵۶/۱۴	۵۸/۸۱	۵۳/۳۹
نوع ۴	۵۶/۴۶	۵۹/۲۵	۵۳/۵۸

اگرچه هدف روش پیشنهادی در این مقاله بهبود کارایی روش پایه VTLN از جهت زمان محاسبات بود، اما نتایج جدول ۳ نشان می‌دهد که علاوه بر این هدف، این روش درصد دقت بازشناسی کلمه را نیز حدود ۰/۵۴ درصد نسبت به سامانه نوع یک بهبود داده است. بررسی تفکیکی این نتایج بر حسب جنسیت گویندگان نشان می‌دهد با توجه به اینکه روش پیشنهادی باعث بهبود دقت بازشناسی نسبت به سامانه نوع یک، بر روی کل داده‌ها شده است، اما برتری این روش از جهت گویندگان زن بوده است. حتی استفاده از مدل رگرسیون جداگانه برای گویندگان مرد نیز این موضوع را تغییر نداده و با وجود بهبود بازشناسی گویندگان مرد در حالت نوع چهار نسبت به نوع سه، همچنین این حالت ضعیف‌تر از نوع یک بوده است. این موضوع را می‌توان به نوع پراکندگی گویندگان مرد ارتباط داد؛ به نحوی که پخش این گویندگان به صورت خطی نبوده و ممکن است در صورت استفاده از مدل‌های با مرتبه بالاتر، نتایج بهتری حاصل شود.

همان‌طور که در شکل (۷) مشخص است، پراکندگی غالب گویندگان مرد و زن در یک راستا نیست. این موضوع نشان می‌دهد که در صورت استفاده از دو مدل خطی جداگانه برای گویندگان مرد و زن، تخمین بهتری از ضریب پیچش به دست خواهد آمد. این موضوع در جدول (۱) با نام نوع چهار بررسی شده است. برای این منظور، برای گویندگانی که امتیاز جنسیت بزرگ‌تر از صفر دارند، یک خط رگرسیون برای گویندگان با ضریب جنسیت کمتر از صفر، خطی دیگر تخمین زده شده است. شکل (۸) خطوط رگرسیون به دست آمده را نشان می‌دهد.



(شکل - ۸): مدل سازی جداگانه توزیع ضریب پیچش بهینه برای گویندگان مرد و زن (نوع ۴). مربع‌های قرمز و دایره‌های آبی رنگ به ترتیب نشان دهنده گویندگان مرد و زن هستند و خط و خط چین سیاه به ترتیب نشان دهنده منحنی به دست آمده از رگرسیون برای هر یک از آنها است.

با توجه به استخراج دو دسته ویژگی متناظر با دو روش تخمین ضریب پیچش، و همچنین حالت پایه (WF=1)، در کل چهار مدل صوتی می‌توان آموزش داد که نتیجه دقت بازشناسی آنها بر روی دادگان آزمون در جدول (۳) آورده شده است.

همان‌طور که مشاهده می‌شود با استفاده از روش پیشنهادی در حالت نوع سه در دادگان آزمون، دقت بازشناسی کلمه کل به حدود ۵۶/۱۴ درصد رسیده است که منجر به افزایش حدود ۱/۶ درصدی نسبت به روش پایه شده است. دقت روش پیشنهادی در حالت نوع چهار، به ۵۶/۴۶ درصد رسیده است که افزایشی در حدود ۱/۹۱ درصد را نسبت به روش پایه نشان می‌دهد. همچنین، این روش سرعت تخمین ضریب پیچش را به میزان قابل توجه ۱۶۹ برابر افزایش داده است (جدول ۲). دقت‌های بازشناسی در

یاسر شکفته، جهان‌شاه کبودیان، محمدمحسن گودرزی و ایمان صراف رضائی، "بهبود کارایی سیستم کاوشگر کلمات تلفنی با استفاده از نرمالیزاسیون امتیاز اطمینان مبتنی بر روش برنامه‌ریزی خطی"، *دوفصلنامه پردازش علائم و داده‌ها*، سال چهارم، شماره پیاپی ۱۴، صص. ۳۷-۴۸، زمستان ۱۳۸۹.

Acero A. and Stern R. 1991, Robust speech recognition by normalization of the acoustic space, in Proc. ICASSP '91, vol. 2, pp. 893-896.

Bijankhan M., Sheykhzadegan J., Roohani M.R., Zarrintare R., Ghasemi S.Z., and Ghasedi M.E. 2003, TFarsDat – The Telephone Farsi Speech Database, In Proc. Eurospeech, Geneva, Switzerland, pp. 1525-1528.

Burget L., Matejka P., and Cernocky, J. 2006, Discriminative training techniques for acoustic language identification. In Proc. ICASSP, vol. 1, pp. I-I.

Claes T., Dologlou I., Bosch L., and Comperolle D. 1998, A novel feature transformation for vocal tract length normalization in automatic speech recognition, IEEE Trans. Speech Audio Process., vol. 11, no. 6, pp. 603-616.

Cohen, J., Kamm, T., and Andreou, A. 1994, An Experiment in systematic speaker variability, Speech Workshop on Robust Speech Recognition.

Cui X. and Alwan A. 2006, Adaptation of children's speech with limited data based on formant-like peak alignment, Comput. Speech Lang., vol. 20, no. 4, pp. 400-419.

Ding G., Zhu Y., Li C., and Xu B. 2002, Implementing vocal tract length normalization in the MLLR framework, in Proc. ICSLP '02, pp. 1389-1392.

Eide E. and Gish H. 1996, A parametric approach to vocal tract length normalization, in Proc. ICASSP '96, pp. 346-349.

Emori T. and Shinoda K. 2001, Rapid vocal tract length normalization using maximum likelihood estimation, in Proc. Eurospeech '01, Aalborg, Denmark.

Farsdet Speech Database, Research Center of Intelligent Signal Processing, <http://www.rcisp.com>

Giuliani D., Gerosa M., and Brugnara F. 2006, Improved automatic speech recognition through speaker normalization, Comput. Speech Lang., vol. 20, no. 1, pp. 107-123.

HTK, HMM ToolKit and HTK book, Available from: <http://htk.eng.com.ac.uk>

Lee L. and Rose R. 1998, A frequency warping approach to speaker normalization, IEEE Trans. Speech Audio Process., vol. 6, no. 1, pp. 49-60.

## ۸- جمع‌بندی

در این مقاله به جبران‌سازی مشکلات مربوط به تفاوت طول مجرای صوتی گویندگان مختلف در حوزه بازشناسی گفتار پرداخته شد. هنجارسازی طول مجرای صوتی به‌عنوان راه حلی مؤثر برای این مشکل بررسی و نحوه اعمال پیچش فرکانسی بر روی طیف گفتار تشریح شد. در ادامه روش متداول به‌دست آوردن ضریب پیچش با استفاده از الگوریتم مبتنی بر جستجو توضیح داده و مشاهده شد که عمده مشکلات این روش زمان‌بر بودن و هزینه محاسباتی بالای آن است. با بررسی توزیع گویندگان مرد و زن بر حسب ضریب پیچش نشان دادیم که ارتباط معنی‌داری بین جنسیت و ضریب پیچش وجود دارد و مبنای روش پیشنهادی را بر همین موضوع استوار کردیم. از این رو سامانه‌ای برای تشخیص جنسیت گوینده ارائه شد که خروجی این سامانه، امتیاز وابسته به جنسیت هر گوینده است؛ سپس با استفاده از امتیاز جنسیتی حاصل و ضرایب پیچش بهینه به‌دست‌آمده از روش جستجو، روشی پیشنهاد شد که در آن ضرایب پیچش از روی امتیاز تشخیص جنسیت با روش کمینه‌کردن خطا، تخمین زده شود. نتایج پیاده‌سازی‌ها کارایی بهتر و سرعت بیشتر این روش را نسبت به روش متداول مبتنی بر الگوریتم جستجو نشان می‌دهد.

## ۹- مراجع

باقر باباعلی، حسین صامتی، هادی ویسی، "بکارگیری نرمالسازی اثر طول مسیر صوتی گوینده‌ها (VTLN) در سیستم بازشناسی گفتار پیوسته فارسی مبتنی بر مدل مخفی مارکوف"، *سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران*، جزیره کیش، اسفند ۱۳۸۶.

شهبلا عزیزی، فرزاد توحیدخواه، فرشاد الماس گنج، "بررسی اثر استفاده از روش تطبیق هنجارسازی طول مسیر صوتی به منظور تشخیص اختلالات گفتاری رایج و گفتاردردمانی کودکان فارسی زبان"، *فصلنامه مهندسی پزشکی زیستی*، سال ششم، شماره ۴، صص. ۲۵۷-۲۶۵، زمستان ۱۳۹۱.

قمرناز تدین تبریزی، سعید ستایشی، "ارائه روشی مبتنی بر نرمالسازی اکوستیکی و خوشه بندی برای بهبود بازشناسی گفتار کودکان فارسی زبان"، *مجله فنی مهندسی دانشگاه آزاد اسلامی مشهد*، صص. ۱۱۳-۱۲۵، زمستان ۱۳۸۸.



Wang S., Cui X., and Alwan A. 2007, Speaker adaptation with limited data using regression-tree-based spectral peak alignment, IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 8, pp. 2454–2464.

Wegmann S., McAllaster D., Orloff J., and Peskin B. 1996, Speaker normalization on conversational telephone speech, In Proc. ICASSP '96, pp. 339–341.

Welling L., Ney H., and Kanthak S. 2002, Speaker Adaptive Modeling by Vocal Tract Normalization, IEEE Trans. Speech and Audio Processing, vol. 10, no. 6, pp. 415-426.

Welling L., Kanthak S. and Ney H. 1999, Improved Methods for Vocal Tract Normalization, In proc. ICASSP, pp. 761-764.

Yoma, N. B., Garretton, C., Huenupan, F., Cetalan, I., and Wuth Sepulveda, J. 2013, On Reducing Harmonic and Sampling Distortion in Vocal Tract Length Normalization, IEEE Trans. Audio, Speech, Lang. Proc., vol. 21, no. 1, pp 110-121.



**یاسر شکفته** تحصیلات خود را در

مقطع کارشناسی در دو رشته مهندسی پزشکی-بیوالکترونیک و مهندسی برق-الکترونیک به ترتیب در سال های ۱۳۸۴ و ۱۳۸۵ در دانشگاه صنعتی امیرکبیر به

پایان رساند. ایشان در سال های ۱۳۸۷ و ۱۳۹۲ مدارک کارشناسی ارشد و دکتری خود را در رشته مهندسی پزشکی (گرایش بیوالکترونیک) از همان دانشگاه اخذ کرد. زمینه های پژوهشی مورد علاقه ایشان پردازش های خطی و غیرخطی سیگنال، شناسایی الگو و پردازش صوت و زبان طبیعی است.

نشانی رایانامه ایشان عبارت است از:

[y\\_shekofteh@aut.ac.ir](mailto:y_shekofteh@aut.ac.ir), [yahoo.com](mailto:y_shekofteh@yahoo.com)



**محمد محسن گودرزی** تحصیلات

خود را در مقطع کارشناسی در دو رشته مهندسی پزشکی-بیوالکترونیک و مهندسی برق-کنترل به ترتیب در سال های ۱۳۸۶ و ۱۳۸۸ در دانشگاه

صنعتی امیرکبیر به پایان رساند. در سال ۱۳۸۹ در مقطع کارشناسی ارشد رشته مهندسی پزشکی-بیوالکترونیک از همان دانشگاه فارغ التحصیل شد. وی هم اکنون در مقطع دکترای مهندسی پزشکی-بیوالکترونیک در دانشگاه صنعتی امیرکبیر در حال تحصیل است. زمینه های پژوهشی مورد

Leggetter C. J. and Woodland P. C. 1995, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Comput. Speech Lang., vol. 9, pp. 171–185.

McDonough J., Schaaf T., and Waibal A. 2004, Speaker adaptation with all-pass transforms, Speech Commun., vol. 41, no. 1, pp. 75–91.

Mimura M. and Kawahara T. 2011, Fast Speaker Normalization and Adaptation based on BIC for Meeting Speech Recognition, IEICE Transaction on information and systems, vol J95-D.

Mitra V., Wang W., Franco H., Lei Y., Bartels C., and Graciarena M. 2014, Evaluating robust features on Deep Neural Networks for speech recognition in noisy and channel mismatched conditions, In proc. Interspeech, pp. 895-899.

Molau S., Kanthak S., and Ney H. 2000. Efficient vocal tract normalization in automatic speech recognition. In Proc. of the ESSV'00.

Panchapagesan S. and Alwan A. 2009, Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC, Comput. Speech Lang., vol. 23, no. 1, pp. 42–46.

Pitz M. and Ney H. 2005, Vocal tract normalization equals linear transformation in cepstral space, IEEE Trans. Speech Audio Process., vol. 13, no. 5, pp. 930–944.

Reynolds, D.A., Quatieri, T.F., and Dunn, R.B. 2000. Speaker verification using adapted Gaussian mixture models. Digital signal processing, 10(1), pp. 19-41.

Sainath T. N., Kingsbury B., Mohamed A., Saon G., and Ramabhadran B. 2014, Improvements to filterbank and delta learning within a deep neural network framework., In proc. ICASSP, pp. 6839-6843.

Sanand D.R., Schlüter R., and Ney H. 2010, Revisiting VTLN using linear transformation on conventional MFCC, In Proc. Interspeech '10, pp. 538–541.

Saon G., Kuo J., Rennie S. and Picheny M. 2015, The IBM 2015 English Conversational Telephone Speech Recognition System, arXiv:1505.05899v1.

Sheikhzadegan J., Bijankhan M. 2006, Persian speech databases. In 2nd Workshop on Persian Language and Computer, pp. 247-261.

Shekofteh Y., Almasganj F. 2013, Autoregressive modeling of speech trajectory transformed to the reconstructed phase space for ASR purposes Digital Signal Processing, vol. 23, pp. 1923–1932

Umesh S., Zolnay A., and Ney H. 2005, Implementing frequency-warping and VTLN through linear transformation of conventional MFCC., In Proc. Interspeech '05, pp. 269–272.



اصلی ایشان پردازش سیگنال و بیش تر در زمینه بازشناسی گفتار فارسی و بازشناسی خصوصیات پروزودیک گفتار است. نشانی رایانامه ایشان عبارت است از:

almas@aut.ac.ir



**شقایق رضا** تحصیلات خود را در مقطع کارشناسی در رشته مهندسی پزشکی (بیوالکترونیک) در دانشگاه صنعتی امیرکبیر (۱۳۸۵) و کارشناسی ارشد را در همان رشته در دانشگاه صنعتی امیرکبیر (۱۳۸۷) به پایان رساند. ایشان هم اکنون دانشجوی مقطع دکتری بیوالکترونیک در دانشگاه صنعتی امیرکبیر هستند. از موضوعات مورد علاقه ایشان می‌توان به پردازش گفتار، پردازش سیگنال و تصویر اشاره کرد. نشانی رایانامه ایشان عبارت است از:

shaghayegh.reza@gmail.com



**ایمان صراف رضایی** مدارک کارشناسی و کارشناسی ارشد خود را در رشته مهندسی پزشکی (گرایش بیوالکترونیک) به ترتیب در سال‌های ۱۳۸۴ و ۱۳۸۷ از دانشگاه صنعتی امیرکبیر اخذ کرده است. وی در حال حاضر مدیر گروه پردازش صوت و زبان طبیعی پژوهشگاه پردازش داده پژوهشگاه توسعه فناوری‌های پیشرفته خواجه‌نصیرالدین طوسی است. مدل‌سازی و پردازش سیگنال گفتار، زمینه عمومی پژوهش‌های ایشان است. نشانی رایانامه ایشان عبارت است از:

imansarraaf@aut.ac.ir

علاقه ایشان شناسایی الگو، پردازش سیگنال‌های تصادفی و پردازش و بازشناسی گفتار است. نشانی رایانامه ایشان عبارت است از:

mm.goodarzi@aut.ac.ir



**حسن قلی‌پور** مدارک کارشناسی و کارشناسی ارشد خود را در رشته مهندسی پزشکی (گرایش بیوالکترونیک) به ترتیب در سال‌های ۱۳۸۹ و ۱۳۹۱ از دانشگاه صنعتی امیرکبیر اخذ کرده است. وی در حال حاضر عضو گروه پردازش صوت و زبان طبیعی پژوهشگاه پردازش داده پژوهشگاه توسعه فناوری‌های پیشرفته خواجه‌نصیرالدین طوسی است. زمینه‌های تخصصی مورد علاقه ایشان پردازش و بازشناسی گفتار و پردازش زبان طبیعی است. نشانی رایانامه ایشان عبارت است از:

h.gholipour@aut.ac.ir



**سیدجهان‌شاه کبودیان** تحصیلات خود را در مقاطع کارشناسی، کارشناسی ارشد و دکترا در دانشکده مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران) گذراند و مدرک دکترای خود را در سال ۱۳۸۹ از دانشگاه مذکور دریافت کرد. وی هم‌اکنون استادیار گروه مهندسی کامپیوتر و فناوری اطلاعات دانشگاه رازی کرمانشاه است. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: پردازش سیگنال، پردازش گفتار و صوت، پردازش زبان طبیعی، شناسایی الگو، یادگیری ماشین و الگوریتم‌های فراابتکاری. نشانی رایانامه ایشان عبارت است از:

kabudian@{razi, aut}.ac.ir



**فرشاد الماس گنج** در سال ۱۳۶۳ در رشته برق، گرایش الکترونیک، از دانشگاه امیرکبیر فارغ التحصیل شد؛ سپس دوره کارشناسی ارشد خود را در همین رشته تا سال ۱۳۶۷ ادامه داد و با یک فاصله چهار ساله، دوره دکترای برق (گرایش مهندسی پزشکی) را آغاز کرد. وی در حال حاضر دانشیار دانشکده مهندسی پزشکی دانشگاه امیرکبیر است. زمینه پژوهشی